2021

Combining assumptions and graphical network into gene expression data analysis

Demba Fofana
The University of Texas Rio Grande Valley, dfofana@yahoo.com

E. O. George

Dale Brown

Recommended Citation

Fofana, D., George, E.O. & Bowman, D. Combining assumptions and graphical network into gene expression data analysis. J Stat Distrib App 8, 9 (2021). https://doi.org/10.1186/s40488-021-00126-z

Fofana *et al. Journal of Statistical Distributions and Applications* (2021) 8:9
https://doi.org/10.1186/s40488-021-00126-z

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Combining assumptions and graphical network into gene expression data analysis

Demba Fofana[1*], E. O. George[2] and Dale Bowman[2]

*Correspondence:
dfofana@yahoo.com
[1]University of Texas Rio Grande Valley, Edinburg, TX 78539, USA
Full list of author information is available at the end of the article

Abstract

Background: Analyzing gene expression data rigorously requires taking assumptions into consideration but also relies on using information about network relations that exist among genes. Combining these different elements cannot only improve statistical power, but also provide a better framework through which gene expression can be properly analyzed.

Material and methods: We propose a novel statistical model that combines assumptions and gene network information into the analysis. Assumptions are important since every test statistic is valid only when required assumptions hold. So, we propose hybrid *p*-values and show that, under the null hypothesis of primary interest, these *p*-values are uniformly distributed. These proposed hybrid *p*-values take assumptions into consideration. We incorporate gene network information into the analysis because neighboring genes share biological functions. This correlation factor is taken into account via similar prior probabilities for neighboring genes.

Results: With a series of simulations our approach is compared with other approaches. Area Under the ROC Curves (AUCs) are constructed to compare the different methodologies; the AUC based on our methodology is larger than others. For regression analysis, AUC from our proposed method contains AUCs of Spearman test and of Pearson test. In addition, true negative rates (TNRs) also known as specificities are higher with our approach than with the other approaches. For two group comparison analysis, for instance, with a sample size of $n = 10$, specificity corresponding to our proposed methodology is 0.716146 and specificities for t-test and rank sum are 0.689223 and 0.69797, respectively. Our method that combines assumptions and network information into the analysis is shown to be more powerful.

Conclusions: These proposed procedures are introduced as a general class of methods that can incorporate procedure-selection, account for multiple-testing, and incorporate graphical network information into the analysis. We obtain very good performance in simulations, and in real data analysis.

Keywords: Bayesian spatial network, Gene expression, Multiple testings

Introduction

x Gene expression data can be analyzed in a multiple testing setting as well as many other statistical methods. The validity of each test depends on the underlying distributional assumptions of the test. A proper analysis of gene expression data requires taking assumptions, usually normality, into consideration (Pounds and Fofana 2012; Pounds and Rai 2009). In addition to incorporating distributional assumptions into the overall testing, it may also be informative to incorporate any prior knowledge of association between entities (Bowman and George 1995). Such associations are often recorded by graphical networks (Wei and Pan 2008). Combining these different elements, besides gaining statistical power, provides a framework through which analysis of gene expression data can be improved. We propose a novel statistical approach that incorporates testing for distributional assumption validity with prior information provided by gene graphical network. In particular, we use graphical networks to incorporate spatial dependence into the analysis of gene expression data. The spatial correlation is taken into account by assuming similar prior probabilities for neighboring genes.

We compare our approach with other methods through a series of simulations, and demonstrate that hybrid-network leads to an improvement on power over other approaches in most of the settings. The comparison of the different methodologies is based on specificities and/or Area under the ROC Curve (AUC). The specificity of a test is called the true negative rate; it is the proportion of samples that test negative using the test in question that are genuinely negative. An ROC curve or a receiver operating characteristic curve shows the performance of a classification model at all classification thresholds. An ROC curve is constructed by reporting sensitivities, true positive rates, on the y-axis and false positive rates on the x-axis.

The network analysis we use is the conditional autoregressive (CAR) model. CAR models are commonly used to represent spatial autocorrelation in data relating to a set of non-overlapping areal units. Those models are typically specified in a hierarchical Bayesian framework, with inference based on Markov Chain Monte Carlo (MCMC) simulation. The most widely used software to fit CAR models is WinBUGS or OpenBUGS. In our work, we use an R function BUGS(·) that helps run OpenBUGS inside R software. Another R function, CARBayes(·), is described in Lee (2013) that can be used for Bayesian spatial modeling with conditional autoregressive priors. Using CARBayes the spatial adjacency information can be specified as a neigbourhood matrix, whereas, with BUGS(·), the user has to specify an adjacency matrix.

Material and methods

Network information can be represented by directed or undirected graphs. Graphs are structures of discrete mathematics and have found applications in scientific disciplines that consider networks of interacting elements, such as genes that interact by sharing some biological resemblances. A graph consists of a set of nodes and a set of edges that connect the nodes. Usually the nodes are the entities of interest. For instance, each gene can be considered a node and the edges the relationships among the genes. A graph can be used in a practical way by developing software to translate between representations, a process sometimes referred to as "coercion".

In data analysis, graphs provide a data structure for knowledge representation, for example in the Gene Ontology (GO). Many studies incorporate gene network information

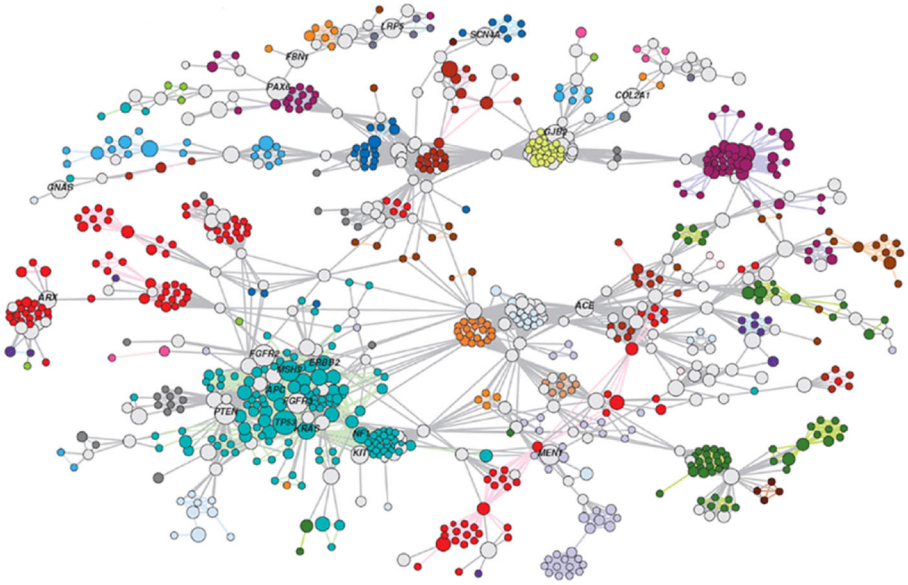

Fig. 1 Undirected Graphs

in data analysis through the GO project. Graphs provide a computational object that can easily and naturally be used to reflect physical objects and relationships of interest. Graphs are important to statistical methodology for exploratory data analysis. A knowledge-representation graph can be juxtaposed with observed data to guide the discovery of important phenomena in the data. In statistical inference, inferential statements about relations between genes due to significantly frequent co-citation, or relation between gene expression and protein complex can be made, (Wei and Pan 2008).

A graph may be directed or undirected. A directed edge is an ordered pair of end-vertices that can be represented graphically as an arrow drawn between the end-vertices. In such an ordered pair the first vertex is called initial vertex or tail and the second the terminal vertex or head. An undirected graph disregards any sense of direction and treats both head and tail identically, see Fig. 1.

Theory/Calculation

Statistical models for hybrid testing

Considerthe following multiple hypothesis testings

$$H_{og} : \boldsymbol{\theta}_g = \boldsymbol{\theta}_{og} \text{ vs } H_{1g} : \boldsymbol{\theta}_g \neq \boldsymbol{\theta}_{og}, g = 1, \cdots, G, \tag{1}$$

with θ_g, a parameter for gene g, and G is the total number of genes, H_{og}, the null hypothesis, and H_{1g} is the alternative hypothesis. Suppose two mutually exclusive test procedures, M_1 and M_2, can be used to perform these statistical tests. When M_1 is used, suppose $\mathbf{T}_1 = \{T_{11}, \cdots, T_{1G}\}$ represents the test statistics and $\mathbf{P}_1 = \{P_{11}, \cdots, P_{1G}\}$ the corresponding set of p-values, and suppose $\mathbf{T}_2 = \{T_{21}, \cdots, T_{2G}\}$ and $\mathbf{P}_2 = \{P_{21}, \cdots, P_{2G}\}$ the corresponding quantities for procedure M_2.

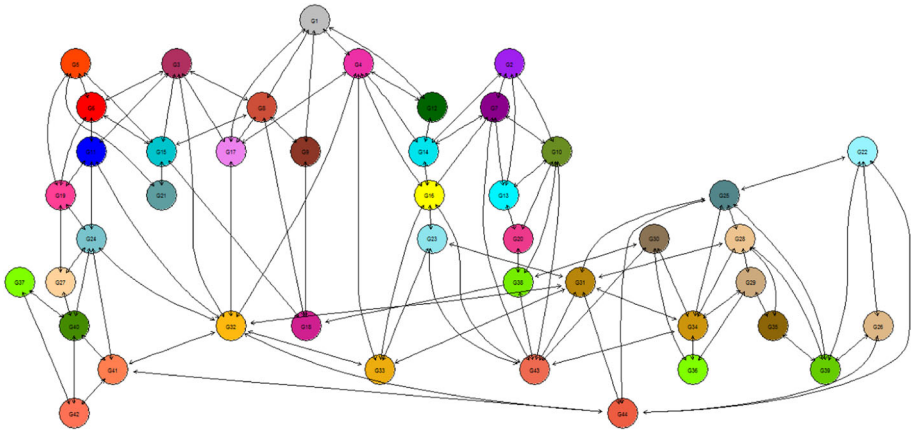

Fig. 2 Simulation Network

Suppose $A_g = i$ is an indication that the assumption for gene g holds for procedure M_i for testing H_{og} vs H_{1g}, $i = 1, 2$. For testing

$$H_{ogA} : A_g = 1,\ g = 1, \cdots, G, \tag{2}$$

suppose $\mathbf{T}_a = \{T_{a1}, \cdots, T_{aG}\}$ are the test statistics obtained from A_g with the corresponding set of p-values $\mathbf{P}_a = \{P_{a1}, \cdots, P_{aG}\}$.
And then, from this method, we define an appropriate summary statistic and denote it by $\mathbf{P} = \{P_1, \cdots, P_G\}$ with

$$P_g = \begin{cases} P_{1g}, & \text{if } A_g = 1 \\ P_{2g}, & \text{if } A_g = 2 \end{cases}$$

$g = 1, \cdots, G$.

The following theorem states the distribution of P_g under the null hypothesis H_{og} of Eq. (1).

Theorem Suppose there are only two mutually exclusive procedures M_1 and M_2 that can be used to test the null hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0. \tag{3}$$

Let P_1 be the p-value obtained if method M_1 is used for testing the null hypothesis H_0, and P_2 be the p-value if method M_2 is used instead. Let P be defined by

$$P = \begin{cases} P_1, & \text{if } M_1 \\ P_2, & \text{if } M_2. \end{cases}$$

Then P is uniformly distributed under the null hypothesis H_0. □

Proof First, we recall some probability theory basics. Let $M_1, M_2, \cdots, M_n$ be a partition of a sample space Ω, that is $M_i \cap M_j = \varnothing\ \forall\ i \neq j$ and $\bigcup_i^n M_i = \Omega$. Then, for any event $E \subset \Omega$,

$$\begin{aligned} E &= E \cap (\textstyle\bigcup_i^n M_i) \\ &= \textstyle\bigcup_i^n (E \cap M_i) \end{aligned}$$

and then for any probability $\mathbb{P}$,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}\left(E \cap \left(\bigcup_i^n M_i\right)\right)\\ \mathbb{P}(E) &= \mathbb{P}\left(\bigcup_i^n (E \cap M_i)\right)\\ &= \sum_i^n \mathbb{P}(E \cap M_i) \text{ (Law of total probability)}\\ &= \sum_i^n \mathbb{P}(E \mid M_i) \times \mathbb{P}(M_i) \text{ (Bayes' rule).}\end{aligned}$$

Also, the law of total probability holds for conditional probability, that is

$$\begin{aligned}\mathbb{P}(E \mid B) &= \sum_i^n \mathbb{P}(E \cap M_i \mid B), \forall \text{ event } B \text{ (Law of total probability)}\\ &= \sum_i^n \mathbb{P}(E \mid M_i, B) \times \mathbb{P}(M_i \mid B) \text{ (Bayes' rule).}\end{aligned}$$

We recall that $\mathbb{P}(\Omega) = \mathbb{P}\left(\left(\bigcup_i^n M_i\right)\right) = \sum_i^n \mathbb{P}(M_i) = 1$.

The question is to show that the hybrid p-value, P, follows a uniform distribution under the null hypothesis (H_0); that is $F_P(p) = \mathbb{P}(P < p \mid H_0) = p, \forall p \in (0,1)$, with F_P the cumulative distribution function of P.

Under the null hypothesis (H_0) of primary interest (gene is expressed, say) and under M_1 and M_2, respectively, both P_1 and P_2 are uniformly distributed, that is $\mathbb{P}(P_1 < p \mid M_1, H_0) = p$ and $\mathbb{P}(P_2 < p \mid M_2, H_0) = p$, see (Pounds and Rai 2009) for instance. Recall that P is a random variable, since P_1 and P_2 are random variables. For the proof, we consider M_1 and M_2 as two events. The notation $\mid H_0$ means under the null hypothesis (H_0).

$$\begin{aligned}\mathbb{P}(P < p \mid H_0) &= \mathbb{P}\left\{(P < p) \cap [M_1 \cup M_2] \mid H_0\right\} \text{ (since } M_1 \text{ and } M_2 \text{ form a partition)}\\ &= \mathbb{P}\left\{(P < p) \cap M_1 \mid H_0\right\} + \mathbb{P}\left\{(P < p) \cap M_2 \mid H_0\right\}\\ &\quad \text{(since } M_1 \text{ and } M_2 \text{ are mutually exclusive)(Law of total probability)}\\ &= \mathbb{P}(P < p \mid M_1, H_0)\mathbb{P}(M_1 \mid H_0) +\\ &\quad \mathbb{P}(P < p \mid M_2, H_0)\mathbb{P}(M_2 \mid H_0)\\ &\quad \text{(Bayes' rule).}\\ &= \mathbb{P}(P_1 < p \mid M_1, H_0)\mathbb{P}(M_1 \mid H_0) + \mathbb{P}\left((P_2 < p) \mid M_2, H_0\right)\mathbb{P}(M_2 \mid H_0)\\ &= p\mathbb{P}(M_1 \mid H_0) + p\mathbb{P}(M_2 \mid H_0)\\ &= p\mathbb{P}(M_1 \mid H_0) + p(1 - \mathbb{P}(M_1 \mid H_0))\\ &= p.\end{aligned}$$

□

Thus P is uniformly distributed under H_0.

Now, transform the p-values by

$$Z_g = \Phi^{-1}\left(1 - P_g\right), \tag{4}$$

where Φ is the cumulative distribution function of the standard normal distribution $N(0,1)$, and P_g is the p-value corresponding to test g. The null distribution of Z_g is the standard normal under H_{og} of Eq. (1). Assume that under the alternative $Z_g \sim N\left(\mu_1, \sigma_1^2\right)$, then

$$f(z_g) = \pi_0 \phi(z_g; 0, 1) + (1 - \pi_0)\, \phi\left(z_g; \mu_1, \sigma_1^2\right), \tag{5}$$

where $\phi(\cdot; \mu_1, \sigma_1^2)$ is the probability density function of $N(\mu_1, \sigma_1^2)$, f is a density function.

Table 1 2—Group simulation specificity comparison

Sample size (n_i)	T-test	Wilcoxon test	Hybrid-Network test
5	0.571726	0.557244	0.575314
10	0.689223	0.69797	0.716146
25	0.884244	0.918197	0.921273
50	0.9839	0.994575	0.994575

Bayesian hierarchical models for spatial data

Conditional autoregressive (CAR) models are commonly used to represent spatial autocorrelation in data relating to a set of non-overlapping areal units. Those data are prevalent in many fields like agriculture (Besag and Higdon 1999), and epidemiology (Lee 2011). There are three different CAR priors commonly used to model spatial autoregression. Each model is a special case of a Gaussian Markov random field (GMRF) that can be written in a general form as

$$\boldsymbol{\phi} \sim N\left(\mathbf{0}, \tau^2 Q^{-1}\right) \tag{6}$$

where Q is a precision matrix that controls for the spatial autocorrelation structure of the random effects, and is based on a non-negative symmetric $G \times G$ neighborhood or weight matrix $\mathbf{W}$, $\mathbf{W} = (w_{kj})$ where $w_{kj} = 1$ if genes k and j are neighboring genes and $w_{kj} = 0$ otherwise, and $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_G)$, is a set of random effects. CAR priors are commonly specified as a set of G univariate fully conditional distributions $\xi(\phi_k \mid \boldsymbol{\phi}_{-k})$ for $k = 1, \cdots, G$ where $\boldsymbol{\phi}_{-k} = (\phi_1, \cdots, \phi_{k-1}, \phi_{k+1}, \cdots, \phi_G)$, and G is the total number of genes (Lee 2013; Lee 2011). The first CAR prior proposed by Besag et al. (1991) is as

$$\phi_k \mid \boldsymbol{\phi}_{-k} \sim N\left(\frac{\sum_{j=1}^{G} w_{kj}\phi_j}{\sum_{j=1}^{G} w_{kj}}, \frac{\tau^2}{\sum_{j=1}^{G} w_{kj}}\right). \tag{7}$$

The conditional expectation is the average of the random effects in neighboring genes, while the conditional variance is inversely proportional to the number of neighbors. The inverse proportionality of conditional variance is due to the fact that if random effects are spatially correlated then the more neighbors a node has the more information there is from its neighbors about the value of its random effect (subject-specific effect). This first CAR prior is used to implement the hybrid-network methodology as in Wei and Pan (2008). The second CAR prior proposed by Leroux et al. (1999) is given by

$$\phi_k \mid \boldsymbol{\phi}_{-k} \sim N\left(\frac{\rho \sum_{j=1}^{G} w_{kj}\phi_j}{\rho \sum_{j=1}^{G} w_{kj} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^{G} w_{kj} + 1 - \rho}\right), \tag{8}$$

while the third CAR prior proposed by Stern and Cressie (1999) is defined by

$$\phi_k \mid \boldsymbol{\phi}_{-k} \sim N\left(\frac{\rho \sum_{j=1}^{G} w_{kj}\phi_j}{\sum_{j=1}^{G} w_{kj}}, \frac{\tau^2}{\sum_{j=1}^{G} w_{kj}}\right), \tag{9}$$

where ρ is a spatial autocorrelation parameter, with $\rho = 0$ corresponding to independence and with $\rho = 1$ corresponding to a strong spatial autocorrelation. A uniform prior on the unit interval is specified for ρ, that is $\rho \sim \cup(0, 1)$, while the usual uniform prior on $(0, M_\tau)$ is assigned to τ^2, with the default value being $M_\tau = 1000$. The intrinsic CAR prior by Besag et al. (1991) is obtained from the second and third CAR priors when $\rho = 1$, while when $\rho = 0$ the difference is on the denominator in the conditional variances.

Table 2 3−Group simulation specificity comparison

Sample size (n_i)	ANOVA test	Kruskal-Wallis test	Hybrid-Network test
5	0.579557	0.57232	0.585729
10	0.668287	0.668287	0.684932
25	0.89141	0.918197	0.929054
50	0.92437	0.9839	0.985663

Standard and spatial normal mixture model

Multipletesting is often an essential step in the analysis of high-dimensional data, such as genomic or proteomic data. The data analysis can be based on p-values, z-scores, t-scores, etc. These test statistics are obtained from data reduction techniques. The hybrid p-values discussed in "Statistical models for hybrid testing" section is an example. Consider for example a test statistic Z. We can assume that across hypotheses $g = 1, \cdots, G$ the test statistic Z_g follows a two-component mixture with density f as in (5). From this two-component mixture two different types of mixture models, the standard and spatial normal mixture models are considered. While spatial normal mixture models consider network information in the analysis, the standard normal mixture models do not.

Standard normal mixture model

In a standard two-component mixture model, Z_g has a density function f of the form

$$f(z_g) = \pi_0 f_o(z_g) + (1 - \pi_0) f_1(z_g), \tag{10}$$

where π_0 is the proportion of genes that are not expressed (null hypothesis), f_o is the distribution of Z_g under the null hypothesis, and f_1 is the distribution of Z_g under the alternative hypothesis.

Spatial normal mixture model

In a spatial normal mixture model, one defines gene-specific prior probabilities

$$\pi_{gs} = \mathbb{P}(T_g = s) \text{ for } g = 1, \cdots, G \text{ and } s = 0, 1, \tag{11}$$

where T_g is defined by

$$T_g = \begin{cases} 1 & \text{if gene g is expressed} \\ 0 & \text{if gene g is not expressed} \end{cases}$$

therefore, the marginal distribution of Z_g is

$$\begin{aligned} f(z_g) &= \textstyle\sum_{s=0}^{1} f\left(z_g \mid T_g = s\right) \mathbb{P}(T_g = s) \\ &= \pi_{g0} f_o(z_g) + \pi_{g1} f_1(z_g), \end{aligned} \tag{12}$$

where z_g is the expression value of gene g for $g = 1, \cdots, G$, and $\pi_{g1} = 1 - \pi_{g0}$. It is believed that genes on the same network, that is a group of genes with the same function, share the same prior probability of expression while different networks have possibly varying prior probabilities. The prior probabilities π_{gs}, based on a gene network, are related to two latent Markov random fields $\mathbf{x}_s = \left\{x_{gs}; g = 1, \cdots, G\right\}, s = 0, 1$ by a logistic transformation:

$$\mathbb{P}(T_g = s) = \pi_{gs} = \frac{\exp(x_{gs})}{\exp(x_{g0}) + \exp(x_{g1})}. \tag{13}$$

Each of the G-dimensional latent vectors $\mathbf{x}_s$ is distributed according to an intrinsic Gaussian conditional auto-regression model (ICAR) (Besag and Kooperberg 1999). The distribution of each spatial latent variable x_{gs} conditional on $x_{-gs} = \{x_{ks}; k \neq g\}$ depends only on its direct neighbors. To be more specific,

$$x_{gs} \mid x_{-gs} \sim N\left(\frac{1}{m_g} \sum_{l \in \delta_g} x_{ls}, \frac{\sigma_{cs}^2}{m_g}\right) \tag{14}$$

where δ_g is the set of indices for the neighbors of gene g, and m_g is the corresponding number of neighbors. The other model specifications are articulated in this way

$$(Z_g \mid T_g = s) \sim N\left(\mu_s, \sigma_s^2\right), \tag{15}$$

$g = 1, \cdots, G$ and $s = 0, 1$. Network structure is summarized in a matrix format called an adjacent matrix: $Adj = (a_{ij})$, $i = 1, \cdots, G; j = 1, \cdots, G$, where

$$a_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and genes i and j are related} \\ 0, & \text{otherwise.} \end{cases}$$

Prior distributions

In a standard normal mixture model, a beta distribution is often assumed as the prior distribution for π_0. In a spatial normal mixture model, gene-specific prior probabilities are introduced. For the spatial normal mixture model, the prior probabilities for π_{gs}, based on a gene network, are related to two latent Markov random fields (MRFs), as mentioned previously. From Eq. (14), we assume priors on the variance components $\sigma_{cs}^2 \sim$ inverse-gamma$(0.01, 0.01)$, the corresponding precision $\frac{1}{\sigma_{cs}^2}$ has gamma$(0.01, 0.01)$ with mean 1 and variance 100. σ_{cs}^2 acts as a smoothing parameter for the spatial field and consequently controls the degree of dependency among the prior probabilities of the genes. The size of σ_{cs}^2 determines how similar the π_{gs} are. The smaller the σ_{cs}^2 are, the more similar the π_{gs}.

Maximum a posterior estimation

A frequentist estimation of a standard mixture model via maximum a posterior estimation (MAPE) is used to show the effectiveness of Bayesian estimation for mixture models. Consider a standard mixture model, Eq. (10), with

$$Z \sim N\left(\mu_s, \sigma_s^2\right) \tag{16}$$

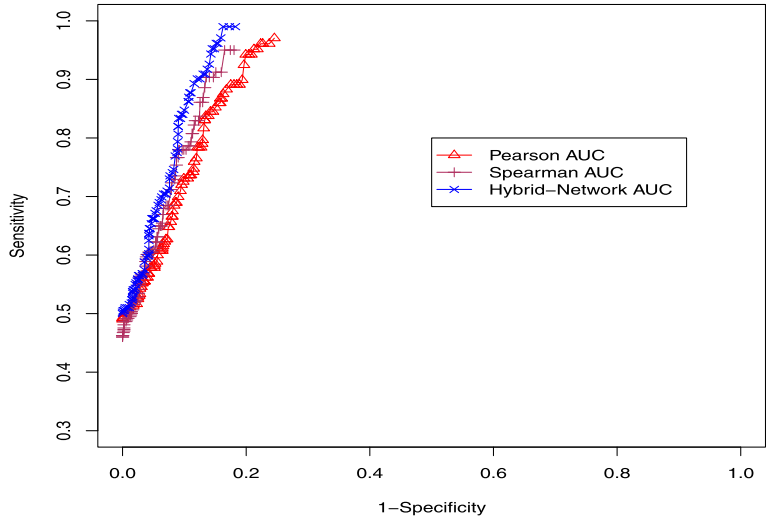

Fig. 3 Comparison of AUC

with $\boldsymbol{\theta}_s = \left(\mu_s, \sigma_s^2\right)$, $s = 0, 1$ and Z is a gene expression test statistic. A direct approach to estimate π_0, π_1, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$ is to compute the likelihood function

$$\begin{aligned} L(\pi_0, \pi_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= \prod_{k=1}^{n} \prod_{g=1}^{G} f(z_{gk}) \\ &= \prod_{k=1}^{n} \prod_{g=1}^{G} \left[\pi_0 f_o(z_{gk}, \boldsymbol{\theta}_0) + \pi_1 f_1(z_{gk}, \boldsymbol{\theta}_1)\right] \end{aligned} \tag{17}$$

and the log likelihood as

$$l(\pi_0, \pi_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \sum_{k=1}^{n} \sum_{g=1}^{G} log[\, \pi_0 f_o(z_{gk}, \boldsymbol{\theta}_0) + \pi_1 f_1(z_{gk}, \boldsymbol{\theta}_1]\,. \tag{18}$$

Obtaining MAPE's of the parameters directly is not possible. To estimate the parameters the expectation-maximization (EM) algorithm may be used. In order to use the EM algorithm, define latent variables $\mathbf{v} = \left\{(v_{gk}, z_{gk}) \mid k = 1, \cdots, n \text{ and } g = 1, \cdots, G\right\}$ where

$$v_{gk} = \begin{cases} 1, & \text{if } g \in \mathbf{G}_1 \\ 0, & \text{if } g \in \mathbf{G}_0 \end{cases}$$

with $\mathbf{G}_0$ (genes not expressed) and $\mathbf{G}_1$ (expressed genes) are null hypothesis and alternative groups respectively, n is the sample common to all genes. If we include latent variables we get complete data, the observed $\mathbf{z}'s$ and the unobserved $\mathbf{v}'s$. The maximum a posterior function for the complete data is

$$L_c(\pi_0, \pi_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \mid \boldsymbol{z}, \boldsymbol{v}) = \prod_{k=1}^{n} \prod_{g=1}^{G} \left[\pi_0 f_o(z_{gk}, \boldsymbol{\theta}_0)\right]^{1-v_{gk}} \left[\pi_1 f_1(z_{gk}, \boldsymbol{\theta}_1)\right]^{v_{gk}}. \tag{19}$$

Taking the log on Eq. (19) we get the log maximum a posterior function as

$$l_c(\pi_0, \pi_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \mid \boldsymbol{z}, \boldsymbol{v}) = \sum_{k=1}^{n} \sum_{g=1}^{G} \left[(1 - v_{gk}) log[\, \pi_0 f_o(z_{gk}, \boldsymbol{\theta}_0)] + v_{gk} log[\, \pi_1 f_1(z_{gk}, \boldsymbol{\theta}_1]\right]. \tag{20}$$

The EM algorithm can be used to obtain MAPE's of π_0, π_1, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, if $(z_{1k}, z_{2k}, \cdots, z_{Gk})$ are assumed to be independent.

However, since there is a graphical network among genes, $(z_{1k}, z_{2k}, \cdots, z_{Gk})$ are not independent. In order to take into account gene graphical network a Bayesian methodology is used. Network analysis is brought into the analysis by generating latent variables according to GMRFs as in Eq. (14). After assigning prior distributions to the parameters, posterior distributions can be found using a partial Gibbs sampler and some Metropolis Hasting algorithms. We use OpenBugs software to get the MAPE's of π_0, π_1, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$.

Statistical inference

The decision rule and acceptance of null hypotheses is based on probabilities from posterior distributions. For each gene g, the point estimate of $p(H_{0g} \mid Data)$ is computed and compared to a threshold τ, for $g = 1, \cdots G$. H_{0g} is rejected when $\hat{p}(H_{0g} \mid Data)$, the point estimate, of $p(H_{0g} \mid Data)$ is less than a threshold τ.

The p-values p_g obtained from the hybrid method are transformed, and the transformed statistics $z_g = \Phi^{-1}(1 - p_g)$ are used, with Φ^{-1} the standard normal quantile function. Through Bayesian modeling, network information is added to the analysis. With the Bayesian inference these posterior estimates are $\hat{\pi}_{g0} = \hat{p}(H_{0g} \mid Data)$. Inferences for the Bayesian hierarchical models are obtained using MCMC simulations, with a combination

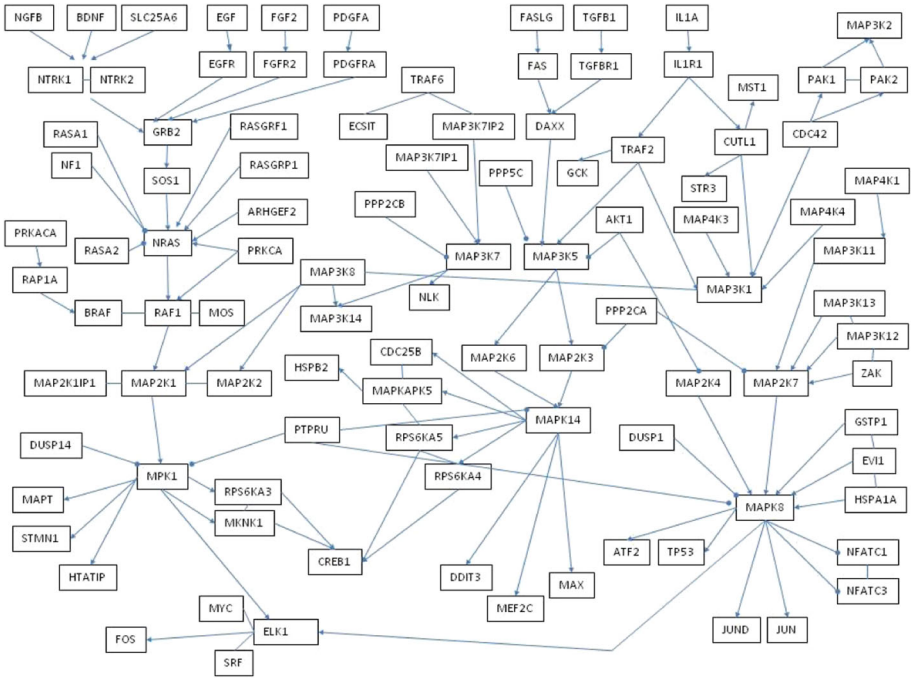

Fig. 4 Gene Graphical Network

of Gibbs sampling and Metropolis steps. Gibbs sampling is used to do MCMC simulation for fully conditional posteriors with closed forms. For those that are not in closed forms the Metropolis-Hasting algorithm is used.

Results

Simulations

To compare the hybrid-network method with other methods, we conducted simulation studies designed to mimic real data analysis. We conducted standard two-group comparison studies (treatment vs control), k-group ($k > 2$) comparison (ANOVA), and regression analysis. The k-group comparison is directly applicable to a genomic study comparing human ependymoma, a brain tumor that occurs in three distinct anatomic regions: Posterior Fossa (PF), Spine (SP), and Supratentorial (ST). Regression analysis is often useful to determine whether, for example, gene expression levels are related to a particular covariate such as DNA synthesis rate (INHIBO).

For each of the three types of analyses conducted in the simulation studies, two different tests can be used. The first one requires the normality assumption while the second may be appropriate when the normality assumption does not hold. For the two-group comparison the hybrid-network method chooses between the standard t-test for normally distributed data and the Wilcoxon test when the normality assumption fails. For k-group ($k > 2$) comparison, the hybrid-network method chooses between the standard ANOVA test and the Kruskal-Wallis test. For the regression analysis, the Pearson test for linear dependency is chosen when the normal assumption holds and the Spearman test if the normality assumption does not hold.

In Eq. (12), we use $\hat{\pi}_{g0}$, the estimate of π_{g0}. And, the decision rule consists of rejecting the null hypothesis, H_{g0}, for gene g, if $\hat{\pi}_{g0}$ is less than a threshold, τ. The conclusion is that the corresponding gene g is expressed. For cancer data analysis, for instance, if a gene is expressed, health researchers will target that gene in finding cure.

The comparison of the different methodologies is mainly based on specificities (not to reject the null hypotheses when they are true, we call them sometimes true negatives). We could provide both specificities and sensitivities (reject the null hypotheses when they are not true, we call them sometimes true positives); but we have decided to compute only specificities because the simulations are computationally intensive.

K-group comparison study

In a group comparison study, gene expression data can be modeled as:

$$Y_{gij} = \mu_g + \tau_{gi} + \epsilon_{gij}, \tag{21}$$

where Y_{gij} is the expression level for gene g of the j^{th} individual in the i^{th} group,

$$g = 1, \cdots, G,\ i = 1, \cdots, k;\ j = 1, \cdots n_i,$$

k is the number of groups, n_i is the sample size of group i, and

$$\epsilon_{gij} \sim N(0,1) \text{ or } \epsilon_{gij} \sim t(\nu),\ \text{ or } \epsilon_{gij} \sim \text{ another distribution.}$$

A 2-group comparison ($k = 2$), interest is in statistical tests of the form

$$H_{g0}: \mu_{g1} = \mu_{g2} \text{ vs } H_{gA}: \mu_{g1} \neq \mu_{g2}, \tag{22}$$

$g = 1, \cdots, G$. Some gene expression levels may be normally distributed while others are not normally distributed. In the two-group comparison study, two tests are often used. The t-test is used when the normality assumption holds and the Wilcoxon test, a non parametric test, is often used when the normality assumption does not hold. For each gene g, a t-test, a Wilcoxon-Mann-Whitney rank sum test, and a Shapiro-Wilk test statistics are computed. Diagnoses for adequacy of the t-test statistics are made through residuals. We compute the residuals from the t-test statistic. We define the residuals on observation, j, in treatment, i, for gene, g, as

$$e_{gij} = Y_{gij} - \hat{Y}_{gij} \tag{23}$$

where $\hat{Y}_{gij}$ is an estimate of the corresponding observation Y_{gij} obtained as follows:

$$\begin{aligned} \hat{Y}_{gij} &= \hat{\mu}_g + \hat{\tau}_{gi} \\ &= \bar{Y}_{g..} + \left(\bar{Y}_{gi.} - \bar{Y}_{g..}\right) \\ &= \bar{Y}_{gi.}. \end{aligned} \tag{24}$$

If the model is adequate, residuals should be structure-less; that is, they should contain no obvious patterns. Through an analysis of residuals, many types of model inadequacies and violations of the underlying assumptions can be discovered. We use the residuals to check for normality. A probit plot of residuals is an extremely useful procedure to test for normality. If the underlying error distribution is normal, this plot will resemble a straight line. Also outliers can be detected through residuals. Outliers show up on probability plots as being very different from the main body of the data. Plotting the residuals in time order of data collection is helpful in detecting correlation between the residuals. This is useful for checking independence assumptions on the errors.

To compare the hybrid-network method with other methods, we perform a simulation study. In this setup, there are two groups of sample sizes varying from 5, 10, 25, and 50. The number of gene expressions having a normal distribution, $N(\mu, 1)$, is 30. For these gene expressions, $\mu = 0$ for the null hypothesis and $\mu = 1$ for the alternative. The remaining gene expressions have log-normal distribution, log-normal $(\mu, 1)$, with $\mu = 0$ in some cases and $\mu = 1$ in other cases. And a graphical network, Fig. 2, is built among genes with 212 number of edges. We translate this graphical network into an adjacent matrix.

The results are presented in Table 1, they show that hybrid-network procedure dominates the other methodologies in most of the settings, since the hybrid-network test specificities are higher than the specificities of the other methods. When the sample size is equal to 5, for instance, the specificity corresponding to the t-test is 0.571726, the specificity corresponding to the Wilcoxon test is 0.557244, and the specificity for the hybrid-network test is 0.575314.

Hybrid ANOVA-Kruskal Wallis study

In a k-group comparison study, a statistical model can be written as Eq. (21). For the model (21), μ_g is a parameter common to all treatments for gene g called the overall mean, and τ_{gi} is a parameter unique to the ith treatment for gene g called the ith treatment effect. Consider the following multiple hypothesis tests

$$H_{g0}\colon \mu_{g1} = \mu_{g2} = \cdots = \mu_{gk} \text{ vs } H_{gA}\colon \mu_{gi} \neq \mu_{gl} \text{ for at least one pair } (i, l) \tag{25}$$

or equivalently, by using the effects models

$$H_{g0}\colon \tau_{g1} = \tau_{g2} = \cdots = \tau_{gk} = 0 \text{ vs } H_{gA}\colon \tau_{gi} \neq 0 \text{ for at least one } i. \tag{26}$$

The hypotheses may be tested using an ANOVA test or the Kruskal-Wallis depending on the normality assumption. If the normality assumption is valid, the ANOVA test is more powerful than the Kruskal-Wallis; and the latter may be more powerful when the normality assumption does not hold. The proposed methodology, hybrid-network, combines a test of assumptions and graphical network information into the analysis. For each gene g, an ANOVA p-value, p_g^a, a Kruskal-Wallis p-value, P_g^w, and a Shapiro-Wilk p-value, P_g^s are computed. We define a hybrid p-value, P_g^h, as

$$P_g^h = \begin{cases} P_g^a, & \text{if } P_g^s \geq \alpha \\ P_g^k, & \text{if } P_g^s < \alpha, \end{cases}$$

for $g = 1, \cdots, G$ where α is a given threshold. The hybrid p-value P_g^h is transformed into a hybrid z-statistic, z_g^h, as follows:

$$z_g^h = \Phi^{-1}\left(1 - P_g^h\right). \tag{27}$$

We use z_g^h to build a CAR model from the given network with the marginal distribution of z_g^h given by

$$f\left(z_g^h\right) = \pi_{g0} f_o\left(z_g^h\right) + \pi_{g1} f_1\left(z_g^h\right), \tag{28}$$

where z_g^h is the expression value for gene g, $g = 1, \cdots, G$.

Table 3 Human ependymoma microarray data

Genes	Gr1	Gr1	···	Gr2	Gr2	···
AKT1	12.48167	11.75317	···	10.95536	11.51737	···
ARHGEF2	14.99632	13.81004	···	13.45263	14.02982	···
ATF2	12.93096	13.14289	···	13.44182	12.72238	···
BDNF	3.392317	4.542258	···	4.716991	5.738768	···
BRAF	9.111918	10.3433	···	10.07682	9.107217	···
CDC25B	10.33114	11.04207	···	11.7139	11.76408	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮

This shows the human ependymoma expression data: genes as gene annotation, groups (Gr1 and Gr2) as sample annotation and real values as gene expression levels.

The prior probabilities π_{gs}, based on a gene network, are related to two latent Markov random fields $\mathbf{x}_s = \{x_{gs}; g = 1, \cdots, G\}, s = 0, 1$ by a logistic transformation:

$$\mathbb{P}(T_g = s) = \pi_{gs} = \frac{\exp(x_{gs})}{\exp(x_{g0}) + \exp(x_{g1})}. \tag{29}$$

The distribution of each spatial latent variable x_{gs} conditional on $x_{-gs} = \{x_{ks}; k \neq g\}$ depends only on its direct neighbors. The proposed CAR prior distribution from (Besag and Kooperberg 1999) is used as

$$x_{gs} \mid x_{-gs} \sim N\left(\frac{1}{m_g}\sum_{l\in\delta_g} x_{ls}, \frac{\sigma_{cs}^2}{m_g}\right), \tag{30}$$

where δ_g is the set of indices for the neighbors of gene g, and m_g is the corresponding number of neighbors.

The hybrid-network methodology, through a series of simulations, is compared to other methods. The setup of these simulations consists of three groups of sample size varying from 5, 10, 25, and 50. The number of genes with the normal distribution $N(\mu, 1)$, $\mu = 0$ for the null hypothesis and $\mu = 1$ for the alternative, is 30. The number of genes with the log-normal distribution, log-normal$(\mu, 1)$, with $\mu = 0$ in some cases and $\mu = 1$ in other cases, is 7 and the number of genes with the Cauchy distribution, Cauchy$(\theta, 1)$, with $\theta = 0$ in some cases and $\theta = 1$ in other cases, is 7. A graphical network is built among genes with 212 edges. We present the simulations results in Table 2. They show that hybrid-network procedure dominates other procedures in most of the cases. When the sample size is 25, for instance, the specificities from the ANOVA test, the Kruskal Wallis and the hybrid-network test are 0.89141, 0.918197, and 0.929054, respectively.

Regression analysis

In microarray regression analysis, a statistical model can be written as

$$Y_{gj} = \beta_{g0} + X_{gj}\beta_{g1} + \epsilon_{gj} \tag{31}$$

where Y_{gj} is the gene expression level for the g^{th} gene in the j^{th} individual with

$$g = 1, \cdots, G,\ j = 1, \cdots, n$$

and some

$$\epsilon_{gj} \sim N(0, 1) \text{ or } \epsilon_{gj} \sim t(\nu), \text{ or } \epsilon_{gj} \sim \text{ another distribution.}$$

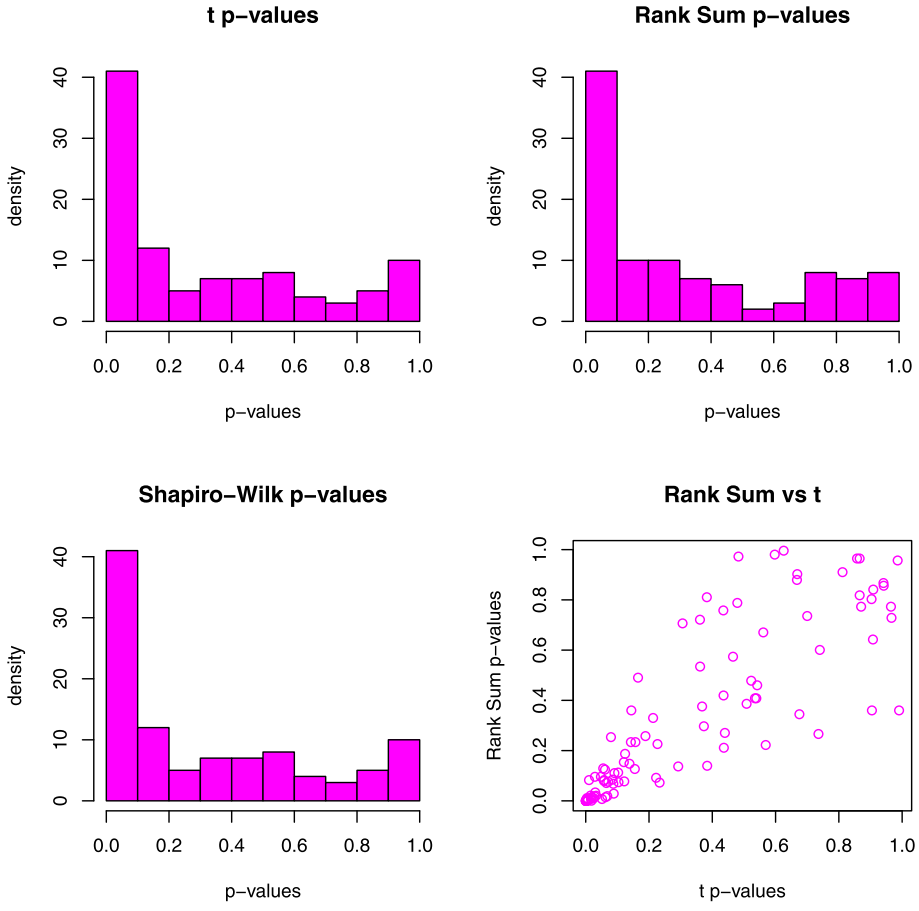

Fig. 5 Tumor Data 2-Group Comparison

The question is whether a response variable and a covariate are correlated. To test for correlation between gene expression with a covariate such as a phenotype, the analysis can be based on Pearson test p-values (P^p), and on Spearman test p-values (P^{sp}). We can use Shapiro-Wilk p-values (P^s) to test for the normality assumptions. Consider, the regression analysis in matrix format

$$\mathbf{Y}_g = \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}_g \tag{32}$$

where

$$\mathbf{Y}_g = \begin{bmatrix} Y_{g1} \\ Y_{g2} \\ \vdots \\ Y_{gn} \end{bmatrix}; \mathbf{X}_g = \begin{bmatrix} 1 & X_{g1} \\ 1 & X_{g2} \\ \vdots & \vdots \\ 1 & X_{gn} \end{bmatrix}; \boldsymbol{\beta}_g = \begin{bmatrix} \beta_{g0} \\ \beta_{g1} \end{bmatrix}; \boldsymbol{\epsilon}_g = \begin{bmatrix} \epsilon_{g1} \\ \epsilon_{g2} \\ \vdots \\ \epsilon_{gn} \end{bmatrix}. \tag{33}$$

We denote the least squares estimators of $\boldsymbol{\beta}_g$ as $\mathbf{b}_g$

$$\mathbf{b}_g = \left(\mathbf{X}_g{}' \mathbf{X}_g\right)^{-1} \mathbf{X}_g{}' \mathbf{Y}_g. \tag{34}$$

Let the vector of the fitted values $\hat{Y}_{gi}$ be denoted as $\hat{\mathbf{Y}}_g$, and the vector of the residual terms $e_{gi} = Y_{gi} - \hat{Y}_{gi}$ be as $\mathbf{e}_g$. The fitted values are represented by

$$\hat{\mathbf{Y}}_g = \mathbf{X}_g \mathbf{b}_g \tag{35}$$

and the residuals by

$$\mathbf{e}_g = \mathbf{Y}_g - \hat{\mathbf{Y}}_g. \tag{36}$$

For each gene g, compute its Pearson p-value, P_g^p, compute its Spearman p-value, P_g^{sp}, and from the residuals from Pearson test, a Shapiro-Wilk test of normality is performed, and for each gene g a p-value, P_g^s, is calculated. Finally, a hybrid p-value, P_g^h is computed as

$$P_g^h = \begin{cases} P_g^p, & \text{if } P_g^s \geq \alpha \\ P_g^{sp}, & \text{if } P_g^s < \alpha \end{cases}$$

where α is a given threshold.

Each hybrid p-value, P_g^h, is transformed into a hybrid z-statistic, z_g^h, as follows:

$$z_g^h = \Phi^{-1}\left(1 - P_g^h\right). \tag{37}$$

Using z_g^h, the marginal distribution of z_g^h is given as

$$f(z_g^h) = \pi_{g0} f_o\left(z_g^h\right) + \pi_{g1} f_1\left(z_g^h\right), \tag{38}$$

where z_g^h is the expression value of gene, g, $g = 1, \cdots, G$.

We compare the hybrid-network with the other procedures through a simulation setup. The setup consists of a sample size of 25. The number of genes with the normal distribution, $N(\mu, 1)$, is 30, $\mu = 0$ for the null hypothesis and $\mu = 1$ for the alternative, and the number of genes with the log-normal distribution, log-normal$(\mu, 1)$, with $\mu = 0$ in some cases and $\mu = 1$ in other cases, is 14. We vary the cutoff point, τ, as in Wei and Pan (2008). And a graphical network is built among genes with 212 number of neighbors. The results of the analysis are presented in Fig. 3. In order to compare the hybrid-testing with other methods, we use AUCs to judge the performance of the proposed method. A greater AUC corresponds to a better methodology. They show that the hybrid-network performs better than the other competing procedures.

Application to human ependymoma microarray

We compare the hybrid-network procedure with the t-test and the Wilcoxon test using human ependymoma data. The data consists of gene expression levels, gene annotation, sample annotation, and a gene graphical network. Figure 4 illustrates a graphical network of the genes under consideration, and Table 3 is a subset of the human ependymoma expression data. In this analysis, there are two groups, the sample sizes are $n_1 = 37$ for group1, $n_2 = 42$ for group2, with the total number of genes of 102, and the number of edges is 196. The data and the R codes can be requested from the corresponding author.

Using Shapiro-Wilk p-values, it appears that some of the expression data are normally distributed and the others are not, with Shapiro-Wilk test p-values less than $\alpha = 5\%$ for some genes. Figure 5 shows histograms of p-values from the t-test, p-values from the rank sum test, and p-values based on the Shapiro-Wilk test of normality, respectively. The last graph of Fig. 5 presents the plot of the p-values from the t-test with respect to

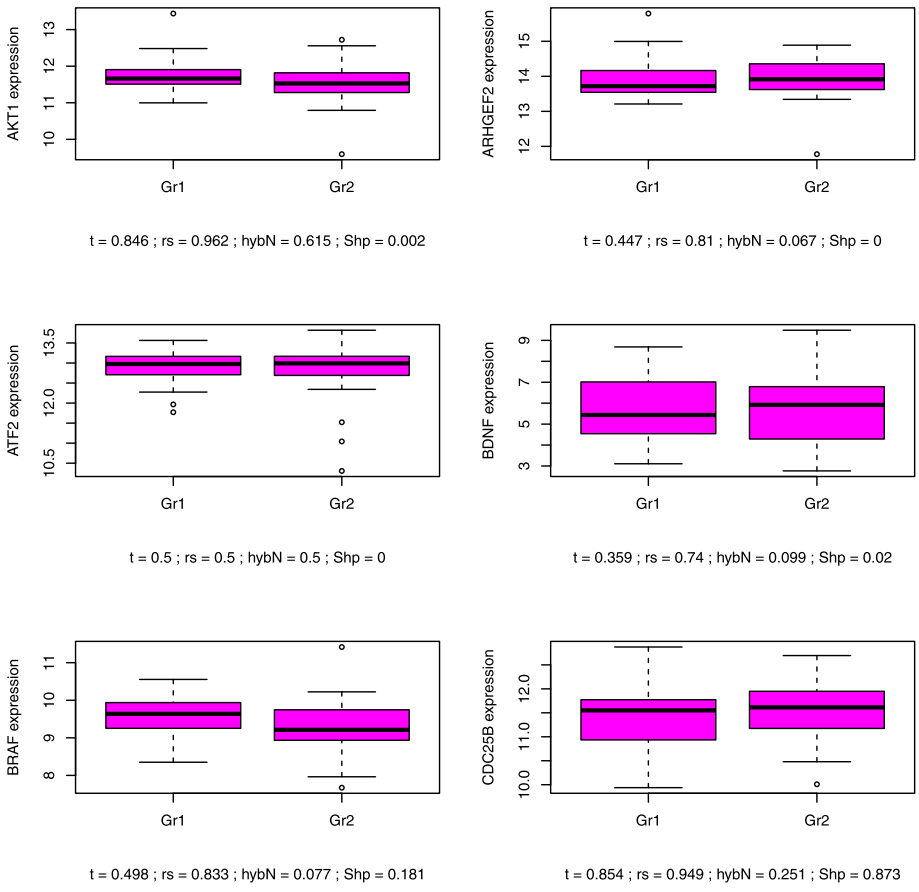

Fig. 6 Tumor Data: Analysis Results

the corresponding p-values from the rank sum test. Using the t-test when the normality assumption is assumed, and the Wilcoxon test otherwise. We apply the hybrid-testing procedure to analyze the data. We incorporate a graphical network to accommodate interactions between genes, as these have been noted to play a crucial role in cell functions (Shojaie and Michailidis 2009).

In order to compare the hybrid-network procedure with the other procedures, we report results for the first six genes. We use box plots as visual methods of comparing groups. Under each Box plot, we report the results, $\hat{\pi}_{.0}$, with t representing the t-test statistic , rs for Wilcoxon test statistic, and $hybN$ for hybrid-network statistic. We also present the p-values from Shapiro Wilk test (Shp) under each box plot. The results are reported on Fig. 6.

With a cutoff point of $\tau = 0.1$, (τ is is a classification threshold, it is like, say α, the level of significance, see Wei and Pan (2008)), all the three methods find that genes $AKT1$, $ATF2$, and $CDC25B$ are not expressed. Only the hybrid-network test finds that the other three genes, $ARHGEF2$, $BDNF$ and $BRAF$ are expressed. This finding is in accordance with the box plot results. The gene selection is based on R head(·) function, that selects the 6-first results. By doing so, we have tried to avoid criticism of biasness in selecting genes to analyze. First, we sort the genes and then pick the 6-first genes for comparing the different methodologies.

Discussion and conclusion

To the best of our knowledge the Hybrid-Network procedure is the very first one that considers assumptions and a graphical network into the analysis of gene expression data. It has a broad variety of applications and entails layers of complexities.

In simulations and in real data analysis, we show that the hybrid-Network procedures perform well. Hybrid-network procedures can be applied to group comparison analysis and to regression analysis. In the near future we are implementing a Hybrid-Network routine that will help researchers analyze gene expressions data in a better and proper manner. In our future research, we plan to apply this method to next generation sequencing data.

Abbreviations

AUC: Area under the ROC curve; ROC: Receiver operating characteristic; GMRF: Gaussian Markov random field

Acknowledgments

This version is presented at the Joint of Statistical Meeting, JSM (The American Statistical Association, Chicago, IL 2016) and is accepted on JSM Proceedings. The authors are highly indebted to participants at this conference for their valuable comments.

Authors' contributions

DF: literature search, model design, simulation, coding, data analysis, data interpretation, writing, critical revision. EOG: model design, writing, critical revision. DB: writing, critical revision, data acquisition. All authors have read and approved the final manuscript.

Funding

The International Conference on Statistical Distributions and Applications (ICOSDA).

Availability of data and materials

Data and coding are available upon request from the corresponding author.

Declarations

Competing interests

The authors declare that they have no competing interests

Author details

[1] University of Texas Rio Grande Valley, Edinburg, TX 78539, USA. [2] University of Memphis, Memphis, TN 38152, USA.

Received: 28 July 2020 Accepted: 25 May 2021
Published online: 08 July 2021

References

Besag, J., Higdon, D.: Bayesian Analysis of Agricultural Field Experients. J. R. Stat. Soc. Ser. B. **61**, 691–746 (1999)
Besag, J., Kooperberg, C.: On conditional and intrinsic autoregressions. Biometrika. **82**, 733–746 (1999)
Besag, J., York, J., Mollié, A.: Bayesian Image Restoration with Two Applications in Spatial Statistics. Ann. Inst. Stat. Math. **43**, 1–59 (1991)
Bowman, D., George, E. O.: Saturated Model for Analyzing Exchangeable Binary Data: Applications to Clinical and Developmental Toxicity Studies. J. Am. Stat. Assoc. **90**, 431 (1995)
Lee, D.: A Comparison of Conditional Autoregressive Models Used in Bayesian Disease Mapping. Spat. Spatiotemporal Epidemiol. **2**, 79–89 (2011)
Lee, D.: CARBayes: An R package for Bayesian Spatial Model with Conditional Autoregressive Priors. J. Stat. Softw. **55**, 13 (2013)
Leroux, B., Lei, X, Breslow, N.: Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. In: Halloran, M. E., Berry, D. (eds.) Statistical Models in Epidemiology, the Environment, and Clinical Trials, pp. 135–178. Springer, New York, (1999)
Pounds, S., Fofana, D.: Hybrid Multiple Testing (2012). http://www.bioconductor.org/packages/2.12/bioc/html/HybridMTest.html. [accessed 12.17.12]
Pounds, S., Rai, S. N.: Assumption adequacy averaging as a concept to develop more robust methods for differential gene expression analysis. Comput. Stat. Data Anal. **53**, 1604-1612 (2009)
Shojaie, A., Michailidis, G.: Analysis of Gene Sets Based on the Underlying Regulatory Network. J. Comput. Biol. **16**, 407–426 (2009)
Stern, H., Cressie, N.: Inference for extremes in disease mapping. In: Lawson, A., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-E., Bertollini, R. (eds.) Disease Mapping and Risk Assessment for Public Health, pp. 63–84. Wiley, Chichester, (1999)
Wei, P., Pan, W.: Incorporating Gene Networks into Statistical Tests for Genomic Data via a Spatially Correlated Mixture Model. Bioinformatics. **24**, 404–411 (2008)

Publisher's Note